Measuring Bias and Agreement in Large Language Model Presupposition Judgments

Katherine Atwell[♭] **Mandy Simons**[⋆] **Malihe Alikhani**[♭]
[♭]Northeastern University [⋆]Carnegie Mellon University
{atwell.ka, m.alikhani}@northeastern.edu
mandysimons@cmu.edu

Abstract

Identifying linguistic bias in text demands the identification not only of explicitly asserted content but also of implicit content including presuppositions. Large language models (LLMs) offer a promising automated approach to detecting presuppositions, yet the extent to which their judgments align with human intuitions remains unexplored. Moreover, LLMs may inadvertently reflect societal biases when identifying presupposed content.

To empirically investigate this, we prompt multiple large language models to evaluate presuppositions across diverse textual domains, drawing from three distinct datasets annotated by human raters. We calculate the agreement between LLMs and human raters, and find several linguistic factors associated with fluctuations in human-model agreement. Our observations reveal discrepancies in human-model alignment, suggesting potential biases in LLMs, notably influenced by gender and political ideology.

1 Introduction

Linguistic acts do not communicate only what is explicitly said, but also implicit assumptions about beliefs and attitudes. Detecting biases embedded in linguistic acts thus requires exploring not only explicit statements but also the underlying **presuppositions**, claims implicitly taken for granted without being directly stated. An example of this phenomenon is the classic example from Russell (1905): the statement "the present king of France is bald" which presupposes that there is a king of France. Recasens et al. (2013), who studied bias mitigation via Wikipedia edits, found that subtle linguistic biases in text often occur via presupposition. Identifying subtle linguistic biases involves recognizing precisely such hidden assumptions embedded beneath explicit assertions. But detecting subtle forms of bias with no clear lexical signals is an ongoing challenge for NLP systems (ElSherief et al., 2021).

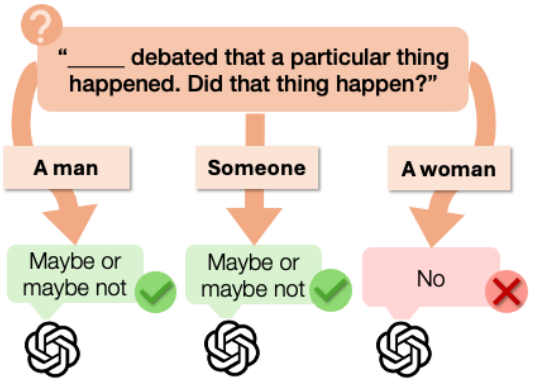

Figure 1: Sample output from GPT illustrating a common trend in our results: when asked about entailment and provided with very little information, substituting a female subject often changed GPT's answer to an incorrect answer.

In this work, we examine whether large language models (LLMs) can be reliably used to identify presupposed content by prompting them to make *projection judgments*, which are commonly used by linguists as a diagnostic tool for presupposed content (§2). To do so, we prompt multiple LLMs to make projection judgments on texts from three English datasets, which contain linguistic presupposition triggers and are annotated with human projection judgments. We calculate agreement scores between humans and LLMs, and utilize NLP tools and existing metadata to determine how factors such as text domain, presupposition trigger, and context impact these agreement scores. Among the factors we study are ones associated with societal biases, such as the gender of the subject and the political ideology of the text.We focus on answering the following research questions:

I. How close are language models' projection judgments to human judgments?
II. What factors impact human-model agreement, and are any of these factors related to societal biases?

Findings of the Association for Computational Linguistics: ACL 2025, pages 2096–2107

We study these questions in detail using three human-annotated datasets that span a variety of text domains and sources, and find that text genre, purported factivity, and trigger type may impact performance. We also find performance gaps when differing political ideologies are discussed. Last, we find substantive evidence that gender may impact LLM performance; consistently lower human-model agreement is observed for projection judgment of clausal complements of verbs when the verb subject signifies a female (Figure 1).

2 Background and Related Work

Presupposition and Projection Presupposition is illustrated by the following simple example:

(1) Sally left the house.

This sentence entails both that Sally was previously in the house, and that she exited the house; speakers of English typically infer both of these entailments. However, when the sentence is negated, the two entailments behave differently:

(2) Sally did not leave the house.

Now, the sentence communicates that Sally did not exit the house; but the implication that Sally was previously in the house remains.

This phenomenon, where entailment survives as an utterance implication under an entailment-canceling operator, is known as **projection**, which is commonly used as a diagnostic for presupposition (Heim, 1983; Van der Sandt, 1992; De Marneffe et al., 2019).

One question being discussed in the current literature on presuppositions concerns the question of how presuppositions are *triggered* (Beaver et al., 2021). There is debate about whether, for instance, clause-embedding verbs can be divided unto two categories: ones that encode presuppositions (factives) and ones that do not (non-factives). The datasets described below provide insight into these questions.

Language models and human projection judgments Several works have used crowd-sourcing to collect human projection judgments with Amazon Mechanical Turk (MTurk) (White and Rawlins, 2018; De Marneffe et al., 2019; Parrish et al., 2021), but only one of these works has studied the impact of various linguistic features on *language model* projection judgments (Parrish et al., 2021), and this work did not study how LLMs behave for this task. This is the first work to comprehensively study LLMs' projection judgments across three different human-annotated datasets, and to closely examine the factors affecting human-model agreement. We study how different linguistic features, such as text genre and trigger type, impact agreement, and examine whether sources of societal bias influence agreement. Below, we describe the three datasets we use to evaluate our baselines in more detail.

3 Datasets

NOPE The NOPE corpus (Parrish et al., 2021) was developed to investigate the context-sensitivity of projection judgments under different presupposition triggers. The authors extracted naturally-occurring sentences from the Corpus of Contemporary American English (COCA) (Davies, 2009) containing ten different types of presupposition triggers. The authors crowd-sourced projection judgments to attain gold labels. They found that transformer models finetuned on these labels exhibited especially high performance on examples with *clefts*, *numeric determiners*, and *temporal adverbs*, and struggled with *implicatives* and *clause-embedding predicates*.

CommitmentBank The CommitmentBank dataset (De Marneffe et al., 2019) was developed to investigate the conditions under which the finite clausal compliments of clause-embedding predicates project (§2). The dataset consists of 1200 naturally-occurring discourse segments from news articles, fiction, and dialogues. Crowd-workers annotated each example based on how certain they believed the speaker was about the truth of the clausal complement (CC), from -3 (certain it is false) to 3 (certain it is true). The authors find that factivity (as standardly assigned in the literature) is a very weak predictor of projectivity. Descriptively, some predicates standardly classified as nonfactive give rise to higher rates of projection than purportedly factive predicates including "know." Overall, results do not support the claim of a categorical factive/nonfactive distinction.

MegaVeridicality The MegaVeridicality dataset (White and Rawlins, 2018) was compiled to test for a correlation between the clause type (declarative or interrogative) required by a verb clause-taking verb V, and two semantic properties: factivity and veridicality (where v is veridical if and only if $x\ v$ that $p \models p$). The authors selected 517 verbs from the MegaAttitude dataset (White and Rawlins,

	NOPE				CommitmentBank				Mega-Veridicality	
model	**Macro F1**				**Spearman**				**Spearman**	
	base	**context**	**FS**	**FS, context**	**base**	**context**	**FS**	**FS, context**	**base**	**FS**
Llama 3.1	0.4888	0.4672	0.4274	0.4247	0.6954	0.7566	0.7583	0.7879	0.5545	**0.8205**
Llama 3.2	0.3028	0.3251	0.3657	0.3513	0.3503	0.4509	0.3374	0.4408	0.0590	0.2716
Llama 3.3	0.4390	0.4346	0.4530	0.4530	0.7109	0.7587	0.7761	**0.7903**	0.5450	0.6355
Mistral	0.4159	0.4199	0.4507	0.4412	0.4590	0.4924	0.6339	0.6730	0.0458	0.2815
Mixtral	0.4431	0.4579	0.4888	**0.5184**	0.6913	0.6046	0.6959	0.6499	0.0689	0.5698
Phi 4	0.4436	0.4884	0.5028	0.5112	0.6444	0.6995	0.7104	0.7456	0.0817	0.0971

Table 1: All macro F1 scores for baselines tested on the NOPE corpus, and all Spearman correlations for the CommitmentBank and MegaVeridicality corpus. All correlations for the CommitmentBank and MegaVeridicality corpus are statistically significant ($p < 0.05$). We experiment with including and excluding context, and few-shot prompting. Our best-performing baseline is few-shot Mixtral with context.

2016) and recruited participants to provide veridicality ratings based on a series of frames such as "Someone {thought, didn't think} that a particular thing happened" and "Someone {was, wasn't} told that a particular thing happened". Raters were asked to answer the question *did that thing happen?* with *yes*, *maybe or maybe not*, or *no*. The authors found that veridicality and factivity do not serve as reliable predictors of clause type.

4 Methods

Prompting Strategies To obtain projection judgments from LLMs, we simulate the human rating tasks used for each dataset (prompt templates in full are provided in Appendix A). For the NOPE dataset, we experiment with prompting for a value (entailment, neutral, or contradiction), as opposed to a number between 0 and 100 (the task given to human annotators), to compare LLM results with the classification model results in the paper.

Experimental Settings We experiment with 6 open-source baselines of different sizes: Llama 3.1:70b, Llama 3.2, Llama 3.3:70b, Mistral, Mixtral, and Phi 4. To limit stochasticity, we set temperature to 0, and because we are prompting the model for short answers, we set max tokens to 5.

5 Results

5.1 Overall Model Performance

In Table 1, we present performance metrics on each dataset. For NOPE, we record macro F1 scores; otherwise, we record Spearman's rank correlations. We provide examples of model responses in Appendix B, and discuss our takeaways below.

Few-shot prompting consistently improves LLM performance We experiment with few-shot prompting strategies and find that in general, models perform better when a few examples are included in the prompt. The most drastic improvements from few-shot prompting are found in the MegaVeridicality corpus, where human-model correlation increases by as much as .5 from the original baseline. We theorize that, because the MegaVeridicality corpus is made up of sentences with low lexical content, models may be hesitant to predict entailment or contradiction unless shown examples.

Adding context to prompts yields mixed results For the NOPE corpus, we find that adding context to our base prompt improves performance for 4 out of 6 baselines, but adding context to our few-shot prompt only yields improvements for 2 out of 6 baselines (though it is worth noting that our best performing baseline is a few-shot baseline with context). For the CommitmentBank dataset, adding context improves results for 5 out of 6 baselines.

Model size does not guarantee improved performance, but it helps for generic statements Phi 4 outperforms Llama 3.3 in all settings on NOPE, despite being a much smaller model. Larger models more consistently outperform smaller models for CommitmentBank, but Mixtral rivals Llama 3.1 for the base prompt. Model size appears to be most impactful for the generic statements in the MegaVeridicality corpus, with consistently large jumps in alignment for larger models.

5.2 Linguistic factors

Fiction and dialogues are associated with higher human-model agreement than news text The CommitmentBank contains texts from Wall Street Journal (WSJ) news articles, British National Corpus (BNC) fiction texts, and Switchboard dialogues. In Table 2, we report the human-model correla-

Model	Domain			Factive		Gender		Ideology - Econ			Ideology - Social		
	WSJ	BNC	SWBD	Yes	No	F	M	Right	Neu.	Left	Right	Neu.	Left
Llama3.1	0.64	0.76	0.76	0.69	0.73	0.67	0.74	0.31^{ns}	0.70	0.45^{ns}	0.97	0.70	0.52
Llama3.2	0.35	0.38	0.17	0.40	0.42	0.35	0.44	-0.07^{ns}	0.41	0.20^{ns}	0.87^{ns}	0.39	0.19^{ns}
Llama3.3	0.61	0.76	0.77	0.67	0.74	0.66	0.75	0.14^{ns}	0.69	0.42^{ns}	0.97	0.68	0.47
Mistral	0.63	0.63	0.61	0.51	0.64	0.59	0.64	0.03^{ns}	0.64	0.68	0.97	0.67	0.52
Mixtral	0.48	0.54	0.73	0.44	0.65	0.45	0.72	-1.00^{ns}	0.45	0.44^{ns}	0.87^{ns}	0.46	0.43^{ns}
Phi 4	0.62	0.70	0.77	0.60	0.69	0.61	0.73	0.60^{ns}	0.64	0.52	0.87^{ns}	0.72	0.48

Table 2: Spearman's rank coefficients indicating human-model agreement for our few-shot baselines on different subsets of the CommitmentBank dataset. All results split by Domain, Factivity, and Gender are statistically significant ($p < 0.05$). We find that domain, factivity, gender, and political ideology are all associated with variations in alignment.

tions for each domain, for our few-shot models with added context. In general, models have higher agreement with human raters for the fiction and dialogue corpora than for news texts, but some exceptions occur (such as Llama 3.2).

LLMs and transformer models differ with respect to the best-performing trigger types Parrish et al. (2021) found that the accuracy of transformers in identifying projective implications depended on trigger type. The best performance was for *clefts*, *numeric determiners*, and *temporal adverbs*, and the worst performance was for *clause-embedding predicates* and *implicatives*. Inversely, we find the highest F1 scores for clause-embedding predicates (Table 3) and models sometimes performed the worst on temporal adverbs. Further, we observe low model performance on clefts.

Purported factives are associated with lower agreement than purported non-factives The CommitmentBank was created to empirically study the purported factive/non-factive distinction. We are interested in studying whether this distinction may impact the relationship between human and model ratings. De Marneffe et al. (2019) find evidence pointing to the absence of a categorical factive/non-factive distinction, as purportedly non-factive predicates are projective to various degrees and some are more projective than purported factives. Here, we study whether factivity, as standardly understood, influences human-model agreement by using the CommitmentBank's labels for purported factivity/non-factivity of each predicate. Across the whole dataset, and within each domain, we calculate the correlations between human and model judgments for factives and non-factives, and report the results in Table 2. We find that all baselines are more aligned with human judgments for non-factives than factives, some by a large margin.

5.3 Social Biases

Human-model agreement on projectivity of clausal complements is consistently lower when sentence subject is female. The CommitmentBank data consists of sentences with main verbs which embed clauses; the question of interest is when and to what degree the content of the embedded clauses is projective. Clause embedding verbs typically describe mental states like belief, or attitudes like regret or being happy, and hence require animate subjects. We use the CommitmentBank metadata and the Gender By Name[1] dataset to study whether subject gender gives rise to differences in human-model agreement. We report the human-model correlations in Table 2, and find that male subjects are associated with a higher correlation than female subjects for all baselines, some by a large margin. This suggests that model inferences may align less with humans' when the subject is female, but the cause is unclear. We find that this pattern is generally consistent across all prompting strategies used (Appendix C).

One possible explanation for this change in agreement is annotator gender bias. To eliminate this potential confounder, we run GPT-3 (text-davinci-003) on the MegaVeridicality dataset and study how human-model agreement changes when the model is presented with indicators of gender as opposed to "someone". Because a portion of the MegaVeridicality dataset denotes its subjects using only the indefinite, genderless pronoun "Someone", it is trivial to change gender of the subject in the prompts. We experiment with substituting "Someone" with "A man" or "A woman" for each example constructed from the [NP _ S] frame. We calculate accuracy and correlation between model predictions on altered examples and average hu-

[1] https://archive.ics.uci.edu/dataset/591/gender+by+name

model	Change of state	Embedded question	Clause embedding predicates	Comparatives	Implicative predicates	Numeric determiners	Reverbs	Aspectual verbs	Temporal adverbs	Clefts
Llama 3.1	0.3839	0.4069	0.5117	0.4111	0.3528	0.4107	0.3773	0.3907	0.3143	0.3171
Llama 3.2	0.3749	0.3122	0.2856	0.2918	0.3278	0.3212	0.3667	0.3582	0.3566	0.3624
Llama 3.3	0.4834	0.4101	0.5333	0.4351	0.3830	0.4151	0.4004	0.4136	0.3756	0.3171
Mistral	0.4245	0.4254	0.5712	0.4781	0.3967	0.3760	0.3220	0.3754	**0.4208**	0.3392
Mixtral	0.3987	**0.4265**	0.5846	**0.7071**	**0.5561**	0.4217	0.3382	0.4725	0.3801	0.4004
Phi 4	**0.5826**	0.4139	**0.6229**	0.4166	0.5053	0.4310	0.3809	**0.5161**	0.3804	0.3695

Table 3: F1 scores for our few-shot baselines with context on different triggers in the NOPE dataset. Counter to the results for transformer models in the original paper, models are often best-performing on clause-embedding predicates.

	Acc.	Pearson	Spearman
Someone	.3647	.4187	.4302
Someone → **a man**	.3040	.1642	.1410
Someone → **a woman**	.2808	.1169	.0981

Table 4: Correlations between model judgments (for GPT-3, $temp = 0$) and human judgments when the prompt given to the model from MegaVeridicality 1) was unchanged, 2) replaced "someone" with "a man", and 3) replaced "someone" with "a woman".

man labels for original example examples using "Someone". As shown in Table 4, differences in agreement were observed for male-gendered vs. female-gendered subjects. We find that the model performs best when given the same prompt as the humans are given, with "Someone" as the subject. When changing the subject to "a man" in the model prompt, we observe a slight drop in accuracy and a larger decrease in correlation. Further, when the subject is changed to "a woman", the accuracy and correlation between model and human ratings drop by several points compared to "a man". These results indicate that human-model agreement is higher when the subject is male rather than female.

A possible explanation is representation; a recent study estimated that texts written by women comprised of around 26.5% of GPT-3's training data (KUNTZ and SILVA, 2023). Another is that these differences are reflections of societal biases in the training data (Kotek et al., 2023). Future work should explore techniques for reducing this gender performance gap; possible strategies that have been successful in existing research include hyperparameter tuning, instruction guiding, and debias tuning (Dong et al., 2024).

The political ideology discussed may impact human-model agreement. We use the set of WSJ articles in the CommitmentBank and run the political ideology classifier developed by (Sinno et al., 2022), which predicts the political ideology under discussion in the text (left, right, or neutral) across three different dimensions: economic, social, and foreign. The average F1 score for this classifier is 0.55. We calculate the correlation between human and model judgments for texts labeled as ideologically left, right, and neutral for the fiscal and social dimensions and compare these correlations. Our results can be found in Table 2. We find that human-model agreement is often lowest for examples labeled as economic right; sometimes, the model judgments are negatively correlated with human judgments. By contrast, for examples labeled socially right, models are more strongly correlated with human judgments than for neutral or left-leaning examples. However, the sample size of texts classified as politically right or left-leaning is comparatively small when compared to those marked neutral; thus, many of the results, particularly for the economic dimension, are statistically insignificant. More research should be done to determine whether these findings are consistent across larger corpora, but it is notable that many of the baselines studied exhibited similar patterns.

6 Conclusion

We provide comprehensive experiments comparing human projection judgments with LLM projection judgments for three distinct, human-labeled datasets. Using six open-source baselines, we find that text domain, trigger type, and factivity can heavily impact human-model alignment. Further, we find evidence that changes to gender and political ideology may impact alignment, suggesting that certain social biases may impact model projection judgments. In particular, we find substantial evidence for disparities due to gender of the subject of clause-embedding verbs.

7 Limitations

Because these datasets were manually annotated, with each example annotated by multiple raters, they are relatively small, on the order of thousands of examples. The set of Wall Street Journal articles in the CommitmentBank is even smaller. Thus, our findings, particularly on bias, should be investigated on a larger scale to determine whether they hold for larger sets across additional text domains.

8 Ethics

In this work, we evaluate the performance of LLMs on existing datasets, and do not release any new publicly-available datasets with gold labels. We also do not use, or release, any LLMs that have previously not been released to the public. We do study the use of LLMs to detect biases that arise from presupposition, and release our prompting techniques for these experiments. However, given that our findings indicate potential biases in LLMs' projection judgments, we urge practitioners to study this technique further before relying on automatic methods alone to detect epistemological biases. If practitioners are to use LLMs to make claims about biases in text, they should also use manual evaluation techniques, and should carefully study the agreement between LLMs and humans, as well as the factors that impact this agreement.

Acknowledgements

References

David I. Beaver, Bart Geurts, and Kristie Denlinger. 2021. Presupposition. In *The Stanford Encyclopedia of Philosophy* (Spring 2021 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

Mark Davies. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics* 14, 2 (2009), 159–190.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, Vol. 23. 107–124.

Xiangjue Dong, Yibo Wang, Philip S Yu, and James Caverlee. 2024. Disclosure and mitigation of gender bias in llms. *arXiv preprint arXiv:2402.11190* (2024).

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 345–363. https://doi.org/10.18653/v1/2021.emnlp-main.29

Irene Heim. 1983. On the projection problem for presuppositions. *Formal semantics–the essential readings* (1983), 249–260.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*. 12–24.

JESSICA B KUNTZ and ELISE C SILVA. 2023. Who Authors the Internet? *Analyzing Gender Diversity in ChatGPT-3 Training Data. Pitt Cyber: University of Pittsburgh* (2023).

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A Corpus of Naturally-Occurring Presuppositions in English. In *Proceedings of the 25th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Online, 349–366. https://doi.org/10.18653/v1/2021.conll-1.28

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 1650–1659. https://aclanthology.org/P13-1162

Bertrand Russell. 1905. On denoting. *Mind* 14, 56 (1905), 479–493.

Barea Sinno, Bernardo Oviedo, Katherine Atwell, Malihe Alikhani, and Junyi Jessy Li. 2022. Political Ideology and Polarization: A Multi-dimensional Approach. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 231–243. https://doi.org/10.18653/v1/2022.naacl-main.17

Rob A Van der Sandt. 1992. Presupposition projection as anaphora resolution. *Journal of semantics* 9, 4 (1992), 333–377.

Aaron Steven White and Kyle Rawlins. 2016. A computational model of S-selection. In *Semantics and linguistic theory*, Vol. 26. 641–663.

Aaron Steven White and Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. In *Proceedings of the 48th annual meeting of the north east linguistic society*, Vol. 3. 221–234.

A Prompt Details

A.1 NOPE

When context is added to the prompts below, it is pre-pended to the target sentence and the result replaces <premise>.

A.1.1 Zero-shot

```
<premise>
Indicate whether you think the
statement is true, false, or
unclear, using the information
in the text above and your
background knowledge of how the
world works.  ONLY respond with
“true", “false", or “unclear".
Do not give any additional text.
Statement:  <hypothesis>
Response:
```

A.1.2 Few-Shot

```
EXAMPLES:
1.  So long as both sides in the
culture war are trapped in the
public sector, there is no escape
from this negative-sum game.
Indicate whether you think the
statement is true, false, or
unclear, using the information
in the text above and your
background knowledge of how the
world works.  ONLY respond with
“true", “false", or "unclear".
Do not give any additional text.
Statement:  There are two sides
in the culture war.
Response:  true
2.  Good morning, junior, welcome
to our country.  Max didn’t
manage to sit up.
Indicate whether you think the
statement is true, false, or
unclear, using the information
in the text above and your
background knowledge of how the
world works.  ONLY respond with
“true", “false", or "unclear".
Do not give any additional text.
Statement:  It would take effort
for Max to sit up.
Response:  unclear
3.  So accounting or computer
science as the main course, with
a little multiculturalism on
the side, would seem to be a far
more useful preparation for the
""culture"" that really matters
to many students today-namely
corporate-bureaucratic
culture-than the study of Homer
or the Bible.
Indicate whether you think the
statement is true, false, or
unclear, using the information
in the text above and your
background knowledge of how the
world works.  ONLY respond with
“true", “false", or "unclear".
Do not give any additional text.
Statement:  The study of Homer
or the Bible is a preparation
for the ""culture"" that really
matters to many students today.
Response:  false

INSTRUCTIONS:<premise>
Indicate whether you think the
statement is true, false, or
unclear, using the information
in the text above and your
background knowledge of how the
world works.  ONLY respond with
“true", “false", or "unclear".
Do not give any additional text.
Statement:  <hypothesis>
Response:
```

A.2 CommitmentBank

In the prompts below, when context is added, <context> is replaced with the given context. When context is not added, <context> is replaced with the empty string.

A.2.1 Zero-shot

```
<speaker_preface><context>
<target>
Tell us how certain <speaker> is
that <prompt>.  Use a scale from
-3 to 3, where -3 means <speaker>
is certain that it is false, 0
means <speaker> is not certain
whether it is true or false, and
3 means <speaker> is certain that
it is true.  ONLY give a number.
```

Do not give any additional text.

A.2.2 Few-shot

EXAMPLES:
1. Speaker: Polly had to think quickly. They were still close enough to shore for him to return her to the police if she admitted she was not an experienced ocean sailor.
Tell us how certain the speaker is that Polly was not an experienced ocean sailor. Use a scale from -3 to 3, where -3 means the speaker is certain that it is false, 0 means the speaker is not certain whether it is true or false, and 3 means the speaker is certain that it is true. ONLY give a number. Do not give any additional text.
Response: 2
2. B: What am I afraid of? A: Yes. B: Um, I don't know if I'm really afraid of spending too much. I just, uh, don't think that I need them, you know.
Tell us how certain speaker B is that she needs them. Use a scale from -3 to 3, where -3 means speaker B is certain that it is false, 0 means speaker B is not certain whether it is true or false, and 3 means speaker B is certain that it is true. ONLY give a number. Do not give any additional text.
Response: -2
3. Speaker: Nick rolled his eyes upwards. "Not so bad, then." She wished she could tell him that Mr Evans hadn't stolen the Will after all but Nick had never thought that he had so there was no point in it.
Tell us how certain the speaker is that Mr Evans hadn't stolen the Will after all. Use a scale from -3 to 3, where -3 means the speaker is certain that it is false, 0 means the speaker is not certain whether it is true or false, and 3 means the speaker is certain that it is true. ONLY give a number. Do not give any additional text.
Response: 1

INSTRUCTIONS:
<speaker_preface><context>
<target>
Tell us how certain <speaker> is that <prompt>. Use a scale from -3 to 3, where -3 means <speaker> is certain that it is false, 0 means <speaker> is not certain whether it is true or false, and 3 means <speaker> is certain that it is true. ONLY give a number. Do not give any additional text.
Response:

A.3 MegaVeridicality

The statements in the MegaVeridicality are intentionally very general, and thus no additional context is provided. Thus, for this dataset we do not experiment with adding context, and only the statement is added in place of <sentence>.

A.3.1 Zero-shot

<sentence>
Did that thing happen? Answer "no", "maybe or maybe not", or "yes". Choose "no" if you are sure that that thing did not happen based on the above statement, and choose "yes" you are sure that that thing did happen based on the above statement. Otherwise, choose "maybe or maybe not". ONLY choose from these three answers and don't provide any other text.
Response:

A.3.2 Few-Shot

EXAMPLES:
1. Someone surmised that a particular thing happened
Did that thing happen? Answer "no", "maybe or maybe not", or "yes". Choose "no" if you are sure that that thing did not happen based on the above

```
statement, and choose "yes"
you are sure that that thing
did happen based on the above
statement.  Otherwise, choose
"maybe or maybe not".  ONLY
choose from these three answers
and don't provide any other text.
Response:  Maybe
2.  A particular person didn't
turn out to have a particular
thing.
Did that thing happen?  Answer
"no", "maybe or maybe not",
or "yes".  Choose "no" if you
are sure that that thing did
not happen based on the above
statement, and choose "yes"
you are sure that that thing
did happen based on the above
statement.  Otherwise, choose
"maybe or maybe not".  ONLY
choose from these three answers
and don't provide any other text.
Response:  No
3.  A particular person was
pained to do a particular thing.
Did that thing happen?  Answer
"no", "maybe or maybe not",
or "yes".  Choose "no" if you
are sure that that thing did
not happen based on the above
statement, and choose "yes"
you are sure that that thing
did happen based on the above
statement.  Otherwise, choose
"maybe or maybe not".  ONLY
choose from these three answers
and don't provide any other text.
Response:  Yes.

INSTRUCTIONS:
<sentence>
Did that thing happen?  Answer
"no", "maybe or maybe not",
or "yes".  Choose "no" if you
are sure that that thing did
not happen based on the above
statement, and choose "yes"
you are sure that that thing
did happen based on the above
statement.  Otherwise, choose
"maybe or maybe not".  ONLY
choose from these three answers
and don't provide any other text.
Response:
```

B Examples of Model/Annotator Agreement or Disagreement

B.1 NOPE Corpus

‘‘*For three nights a comet flared through the desert sky. The winds hooted like owls. A red smudge appeared on the moon."* This text was labeled as entailing the following statement in the NOPE corpus: *A red smudge couldn't be seen on the moon before.* The negation (*"A red smudge appeared on the moon."*) was also labeled as entailing this statement. However, although Llama correctly predicted that the negation entailed that *"a red smudge couldn't be seen on the moon before"*, the model incorrectly chose "unclear" for the original statement, indicating the **Neutral** label. This represents a case where the LLM was able to correctly judge entailment for a statement's negation, but not for the original statement. Thus, the LLM incorrectly predicts that projection does not occur in this case.

B.2 CommitmentBank Dataset

Agreement *"Claudia could see that locking up a Masai for a crime he did not understand was cruel and inhuman."* For this example, all annotators labeled this statement a 3, indicating agreement that this statement definitely entailed the following: "locking up a Masai for a crime he did not understand was cruel and inhuman." Llama 3.1:70b also marked this example as a 3.

Disagreement From an inspection of the data, we find that cases where Llama 3.1:70b disagrees most with the average annotator score occur for the CommitmentBank when the annotators also disagree with one another. An example is "She wished she could say she was sorry now not in the middle of the night when he was asleep." The annotators disagree whether this entails that "she was sorry now", with the following score distribution: 1, 2, -3, -1, 1, 2, 0, 3, 3. In this instance, the LLM agrees with the annotators who marked 3 (indicating entailment), which marks a departure from the average annotator score of 0.89.

C CommitmentBank Gender and Political Ideology Results for All Baselines

Here, we report the full CommitmentBank results on different genders and political ideologies, for all four of our tested baselines (zero- and few-shot, with- and without-context). We report these results on Table 5 and discuss whether any of these prompting techniques appear to mitigate or increase performance gaps between genders and political ideologies.

Gender We find that performance discrepancies by gender persist for all baselines, with models consistently performing better on male names and pronouns than female names and pronouns. Adding context appears to reduce discrepancies for Mistral and Mixtral in the zero-shot case, but does not show a consistent pattern for Llama models. By contrast, adding context often widens gender performance gaps for few-shot prompting. No consistent patterns appear regarding the gender performance gap for few-shot prompting vs. zero-shot prompting approaches.

Political ideology We find that adding context to zero-shot and few-shot prompts often results in improved scores for left-leaning ideologies for Llama models. Other than that, few consistent patterns emerge when comparing baselines, besides the same general trend where models often perform the worst when right-leaning ideologies are discussed and best when center-leaning ideologies are discussed.

Baseline, no context								
Model	**Gender**		**Ideology - Econ**			**Ideology - Social**		
	Female	Male	Right	Center	Left	Right	Center	Left
Llama 3.1	0.64	0.64	0.27^{ns}	0.69	0.29^{ns}	0.89	0.67	0.38
Llama 3.2	0.21	0.36	0.34^{ns}	0.46	0.11^{ns}	0.71^{ns}	0.48	0.10^{ns}
Llama 3.3	0.62	0.66	0.42^{ns}	0.66	0.35^{ns}	0.89	0.65	0.42
Mistral	0.39	0.42	0.20^{ns}	0.38	0.57	0.45^{ns}	0.43	0.29
Mixtral	0.51	0.67	–	0.49	0.73	0.87^{ns}	0.69	0.42^{ns}
Phi 4	0.63	0.59	-0.14^{ns}	0.69	0.34^{ns}	0.87^{ns}	0.68	0.39
Baseline, with context								
Model	**Gender**		**Ideology - Econ**			**Ideology - Social**		
	Female	Male	Right	Center	Left	Right	Center	Left
Llama 3.1	0.70	0.73	0.27^{ns}	0.72	0.34^{ns}	0.89	0.69	0.53
Llama 3.2	0.32	0.47	0.62^{ns}	0.40	0.32^{ns}	0.58^{ns}	0.44	0.18^{ns}
Llama 3.3	0.68	0.75	0.42^{ns}	0.74	0.25^{ns}	0.89	0.71	0.54
Mistral	0.48	0.47	-0.21^{ns}	0.32	0.10^{ns}	0.32^{ns}	0.25	0.28^{ns}
Mixtral	0.53	0.56	1.00^{ns}	0.70	0.59^{ns}	1.00^{ns}	0.72	0.64
Phi 4	0.62	0.66	0.06^{ns}	0.63	0.36^{ns}	0.87^{ns}	0.57	0.52
Few-Shot, no context								
Model	**Gender**		**Ideology - Econ**			**Ideology - Social**		
	Female	Male	Right	Center	Left	Right	Center	Left
Llama 3.1	0.62	0.70	0.43^{ns}	0.65	0.35^{ns}	0.95	0.68	0.38
Llama 3.2	0.29	0.32	0.24^{ns}	0.40	0.08^{ns}	0.87^{ns}	0.38	0.30^{ns}
Llama 3.3	0.64	0.72	0.35^{ns}	0.67	0.30^{ns}	0.97	0.66	0.42
Mistral	0.56	0.61	0.39	0.67	0.22	0.29	0.71	0.29
Mixtral	0.61	0.73	–	0.65	0.49^{ns}	0.11^{ns}	0.72	0.42^{ns}
Phi 4	0.63	0.68	0.33^{ns}	0.69	0.63	0.87^{ns}	0.74	0.52
Few-Shot, with context								
Model	**Gender**		**Ideology - Econ**			**Ideology - Social**		
	Female	Male	Right	Center	Left	Right	Center	Left
Llama 3.1	0.67	0.74	0.31^{ns}	0.70	0.45^{ns}	0.97	0.70	0.52
Llama 3.2	0.35	0.44	-0.07^{ns}	0.41	0.20^{ns}	0.87^{ns}	0.39	0.19^{ns}
Llama 3.3	0.66	0.75	0.14^{ns}	0.69	0.42^{ns}	0.97	0.68	0.47
Mistral	0.59	0.64	0.03^{ns}	0.64	0.68	0.97	0.67	0.52
Mixtral	0.45	0.72	-1.00^{ns}	0.45	0.44^{ns}	0.87^{ns}	0.46	0.43^{ns}
Phi 4	0.61	0.73	0.60^{ns}	0.64	0.52	0.87^{ns}	0.72	0.48

Table 5: All Spearman's rank correlations for the base and few-shot prompts with and without context, disaggregated by gender and political ideology. All correlations for female and male subjects are statistically significant ($p < 0.05$) and all correlations for the center ideology are statistically significant ($p < 0.05$), for both economic and social dimensions. The two results marked with dashes indicate that too few valid responses were generated by the LLM to calculate a correlation; both of these instances occurred with Mixtral.